COMPUTER GENERATION OF DATA SETS FOR HOMEWORK EXERCISES IN SIMPLE REGRESSION

by

BU-655-M September, 1978

S. R. Searle and P. A. Firey

Biometrics Unit, Cornell University, Ithaca, New York

Abstract

A method is given for generating different data sets for homework exercises in simple regression, such that not only are the estimated slopes, intercepts and correlations the same in all data sets, but so also are the analysis of variance tables, with error mean square that is a perfect square. An example is given for $n = 17$, $b = 2$, $a = 6$, $r = .25$ and $\hat{\sigma}^2 = 180^2$. A computer program REGDATA is available for generating data sets of this nature.

1. Introduction

The tedious labor of the arithmetic required in statistical analyses is readily avoidable today by making use of computer program packages or pre-programmed pocket and desk calculators. This is especially true of regression analysis. Nevertheless, in the early stages of learning statistics it is particularly instructive for students, at least once in their lives, to do all the basic arithmetic of an analysis. Furthermore, giving students their own, individually different data sets motivates them to do all their own calculations; in contrast, giving them all the same data set readily allows plagiarization. But arbitrary data sets, for which computed analyses are all different, greatly add to the instructor's burden of

---

Paper No. BU-655-M in the Biometrics Unit Mimeo Series.

grading, especially for large classes. To alleviate this problem in the case of simple regression, we here describe a method for generating numerous different data sets for homework exercises, in which all the computed analyses have the same pre-assigned values of the regression slope b, of the intercept a, of the product-moment correlation r, and of the analysis of variance for fitting the regression, with the error mean square $\hat{\sigma}^2$ being the square of an integer. An example is given for n = 17 data points having b = 2, a = 6, r = .25, and $\hat{\sigma}^2 = 180^2$. A computer program REGDATA is available from the authors for generating data sets of this nature.

Edwards [1959] describes a method for constructing data pairs (x_i, y_i) for $i = 1, \cdots, N$, in such a way that any three of the values

$$b = S_{xy}/S_x^2, \qquad r = bS_x/S_y \tag{1}$$

$$S_x^2 = \sum_i (x_i - \bar{x})^2 \quad \text{and} \quad S_y^2 = \sum_i (y_i - \bar{y})^2 \tag{2}$$

can be chosen arbitrarily where

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) \tag{3}$$

and where $\bar{x}$ and $\bar{y}$ are the arithmetic averages of the N x_i's and y_i's. The method depends on choosing two arbitrary series of positive and negative integers, x_i and e_i for $i = 1, \cdots, N$, in such a way that

$$\bar{x} = 0, \quad \bar{e} = 0, \quad \text{and} \quad \sum_i x_i e_i = 0 . \tag{4}$$

Then y_i is developed as

$$y_i = x_i + e_i \quad \text{with} \quad S_y^2 = S_x^2 + S_e^2 \tag{5}$$

where $S_e^2 = \sum_i (e_i - \bar{e})^2$. This is done by first assigning values to r and N, which

we denote by n - 1, i.e.,

$$N = n - 1 . \tag{6}$$

Values for x_i and e_i are then chosen in such a way that

$$S_x^2 = (n - 1)r^2D^2, \quad S_e^2 = (n - 1)(1 - r^2)D^2 \quad \text{and} \quad S_y^2 = (n - 1)D^2 \tag{7}$$

where D is some multiple of ten chosen so as to eliminate decimals from (7).

We extend this method first by constructing n = N + 1 data pairs, with n - 1 = N of them chosen exactly as described by Edwards [1959] with the n'th one being $x_n = 0 = y_n$. This in no way affects the calculation of (1) through (4). Our further extension is based on the analysis of variance shown in Table 1, for fitting the simple regression of y_i on x_i for i = 1, ···, n.

Table 1: Analysis of variance for fitting the regression of y_i on x_i, i = 1, ···, n.

Source of variation	Degrees of freedom	Sum of Squares	
		Calculated	Assigned[1/]
Mean	1	$n\bar{y}^2$	$= n(a + bm)^2$
Regression	1	bS_{xy}	$= (n-1)r^2k^2b^2D^2$
Residual	n - 2	$S_y^2 - bS_{xy}$	$= (n-1)(1-r^2)k^2b^2D^2$
Total	n	$\sum_{i=1}^{n} y_i^2$	$= (n-1)k^2b^2D^2 + n(a + bm)^2$

[1/] Described in Section 2.

Using (1), the residual mean square in Table 1 is

$$\hat{\sigma}^2 = (S_y^2 - bS_{xy})/(n - 2) = (1 - r^2)S_y^2/(n - 2) . \tag{8}$$

We generate data sets such that not only do b and r of (1) have pre-assigned values

but also $\hat{\sigma}^2$ of (8) is a perfect square. With $S_y^2 = (n - 1)D^2$ from (7), this entails choosing n, r and D so that

$$\frac{(n-1)(1-r^2)D^2}{n - 2} = (\text{integer})^2 \ . \tag{9}$$

Finally, since the discussion up to this point is based on (4), we modify the x_i's and y_i's to have non-zero means, so that the analysis of variance in Table 1 is appropriate.

2. Method

Data sets are generated by the Edwards' method using values of r and n that satisfy (9). Since the data sets are to be for homework exercises, for which students will be doing all the calculations including sums of squares, n must not be large: n of the order of 20 seems appropriate. Also, the Edwards' method requires N = n - 1 to be even. With these limitations we find n = 17 and r = 0.25 are convenient values satisfying (9). We also choose D = 30 so that (6) gives

$$S_x^2 = 900, \quad S_e^2 = 13,500 \quad \text{and} \quad S_y^2 = 14,400 \ . \tag{10}$$

To facilitate making use of Edwards' method for N = n - 1 = 16, we have developed a table of sets of four positive integers in increasing order of their sums of squares t = 4, ···, 400, so that s_t represents four integers n_1, n_2, n_3 and n_4 having $n_1^2 + n_2^2 + n_3^2 + n_4^2 = t$. Not all possible sets s_t are included in the table; for each t, it contains only those for which all $n_i \leq 20$ and for which $t - \frac{1}{4}(\Sigma n_i)^2$ is a minimum. This is done so as to lessen the occurrence of "outliers" in the generated data sets. Based on this table, our method of generation is as follows.

1. Randomly generate four positive integers with sum of squares s^*_t equal to t in the range 50 - 400. Define these integers and those in the s-table corresponding to s_{450-t} as $x_1, x_2, \cdots, x_8$.

2. Define $x_{i+8} = -x_i$ for $i = 1, \cdots, 8$. We then have $\bar{x} = 0$ and $S_x^2 = 900$.

3. Randomly generate four positive integers with sum of squares s^*_t equal to t in the range 50 - 200. Define these integers and those in the s-table corresponding to s_{270-t} as $e_1, e_2, \cdots, e_8$.

The ranges placed on t in steps 1 and 3 are arbitrary, and designed to lessen the occurrence of "outliers".

4. If $\sum_{i=1}^{8} e_i$ is not an even number, repeat step 3 until it is.

5. By exhaustive search, find 8 values p_i, each of them either +1 or -1, such that $\Sigma p_i e_i = 0$. If the e_i's are such that this cannot be done, repeat steps 3 and 4 until it can.

6. Replace e_i by $5p_i e_i$, and then define $e_{i+8} = e_i$ for $i = 1, \cdots, 8$. (The factor of 5 is used here solely to simplify step 3.) We then have

$$\bar{e} = 0, \quad S_e^2 = 13,500, \quad S_{ex} = 0, \quad S_{x(e+x)}/S_x^2 = 1$$

and

$$S_{x(e+x)}/\sqrt{S_x^2 S_{e+x}^2} = 0.25.$$

7. For pre-assigned values of the intercept a and the regression slope b, and for arbitrary constants k and n, for $i = 1, \cdots, 16$

$$\text{replace } x_i \text{ by } kx_i + m,$$
$$\text{replace } y_i \text{ by } kb(x_i + e_i) + (a + bm)$$

and define

$$x_{17} = m$$
$$y_{17} = a + bm.$$

Often, it may be convenient to choose a, b, k and m so that all the x_i and y_i are positive.

8. Re-arrange the data points (x_i, y_i) in a different sequence so as to disguise any pattern in them that may be apparent from their mode of construction.

The n = 17 data points (x_i, y_i) for i = 1, ···, 17 derived in this manner will have the following characteristics:

$$
\begin{array}{lll}
\text{regression intercept} & = a & \bar{x} = m \\
\text{regression slope} & = b & \bar{y} = a + bm \\
\text{correlation} = r & = 0.25 &
\end{array}
\qquad (11)
$$

$$
\begin{aligned}
S_x^2 &= k^2D^2 && = 900k^2 \\
S_y^2 &= (n-1)k^2b^2D^2 && = 14,400k^2b^2 \\
S_{xy} &= (n-1)r^2k^2bD^2 && = 900k^2b \\
\hat{\sigma}^2 &= \frac{(n-1)(1-r^2)k^2b^2D^2}{n-2} && = 900k^2b^2
\end{aligned}
$$

analysis of variance: as in Table 1.

3. Example

Our example for D = 30, n = 17 and r = .25 has a = 6, b = 2, k = 3 and m = 37, and is shown in Table 2.

Table 2: Example of Method.

i	Step 1		Step 2	Steps 3, 4		Step 5	Step 6	Step 7 [1/]	
	x_i		x_i	e_i		$p_i e_i$	$5p_i e_i \to e_i$	x_i	y_i
1	5	s^*_{222}	5	4	s^*_{157}	4	20	52	230
2	4		4	5		-5	-25	49	-46
3	10		10	4		4	20	67	260
4	9		9	10		10	50	64	434
5	10	s_{228}	10	4	s_{113}	4	20	67	260
6	8		8	6		-6	-30	61	-52
7	8		8	6		-6	-30	61	-52
8	0		0	5		-5	-25	37	-70
9			-5				20	22	170
10			-4				-25	25	-94
11			-10				20	7	140
12			-9				50	10	326
13			-10				20	7	140
14			-8				-30	13	-148
15			-8				-30	13	-148
16			0				-25	37	-70
17								37	80

1/ $kx_i + m \to x_i$

$kb(x_i + e_i) + (a + bm) \to y_i$ } for $a = 6$ $b = 2$ $k = 3$ $m = 37$

Then, in accord with Table 1 and results (11) the data points (x_i, y_i) of Step 7 in Table 2 have the following characteristics:

Table 3: Analysis of Variance of Data in Table 2.

Source of variation	Degrees of freedom	Sum of Squares
Mean	1	108,800
Regression	1	32,400
Residual	15	486,000
Total	17	627,200

$a = b$ $\quad n = 17$ $\quad S_x^2 = 8,100$

$b = 2$ $\quad \bar{x} = 37$ $\quad S_y^2 = 518,400$ $\quad \hat{\sigma}^2 = 180^2$

$r = \frac{1}{4}$ $\quad \bar{y} = 80$ $\quad S_{xy} = 16,200$

Reference

Edwards, Bruce [1959]. Constructing simple correlation problems with predetermined answers. American Statistician 13, No. 5, pp. 25-27.